Published in final edited form as:
Int J Data Min Bioinform. 2008 ; 2(2): 176–192.

Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra

Hyun-Woo Cho,
Department of Industrial and Information Engineering, University of Tennessee, Knoxville, TN 37996, USA

Seoung Bum Kim*,
Department of Industrial and Manufacturing Systems Engineering, University of Texas at Arlington, Arlington, TX 76019, USA

Myong K. Jeong,
Department of Industrial and Information Engineering, University of Tennessee, Knoxville, TN 37996, USA

Youngja Park,
Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, GA 30322, USA

Nana Gletsu,
Department of Surgery, Emory University, Atlanta, GA 30322, USA

Thomas R. Ziegler, and
Clinical Biomarkers Laboratory, Center for Clinical and Molecular Nutrition, Department of Medicine, Emory University, Atlanta, GA 30322, USA

Dean P. Jones
Clinical Biomarkers Laboratory, Center for Clinical and Molecular Nutrition, Department of Medicine, Emory University, Atlanta, GA 30322, USA

Hyun-Woo Cho: hcho7@utk.edu; Myong K. Jeong: mjeong@utk.edu; Youngja Park: medyp@emory.edu; Nana Gletsu: ngletsu@emory.edu; Thomas R. Ziegler: tzieg01@emory.edu; Dean P. Jones: dpjones@emory.edu

Abstract

High-resolution Nuclear Magnetic Resonance (NMR) spectroscopy in combination with multivariate statistical methods has been widely used to investigate metabolic fluctuations in biological systems. This study presents three feature selection methods for identifying the metabolite features that contribute to the distinction of spectral samples among varying nutritional conditions in human plasma. Loading vectors of Principal Component Analysis (PCA), the optimal discriminant direction of Fisher discriminant analysis, and index values of the Variable Importance in Projection (VIP) in a Partial Least Square Discriminant Analysis (PLS-DA) were used to calculate the importance of individual metabolite feature in spectra. In addition, an Orthogonal Signal Correction (OSC) filter was used to eliminate unnecessary variations in NMR spectra and its effectiveness was demonstrated through PCA and kernel PCA. For the evaluation of presented feature selection methods, we compared the ability of classification based on the metabolite features selected by each method. The results have shown that the best classification was achieved using VIP values from an OSC-PLS-DA model.

*Corresponding author: sbkim@uta.edu.

Keywords

Nuclear Magnetic Resonance; NMR; feature selection; metabolomics; multivariate statistical analysis; Orthogonal Signal Correction; OSC

1 Introduction

Development of advanced sensing technology has multiplied the sheer volume of spectral data, which is one of the most common types of data in many research fields to which multivariate statistical methods are applied. Examples of spectral data include Near-Infrared (NIR), Mass Spectroscopy (MS), and Nuclear Magnetic Resonance (NMR) spectroscopy. These spectral data increasingly are being used to determine concentrations of samples and to infer other useful properties as a means to uncover patterns inherent in information-rich data (Sun, 1997; Qin, 2003).

Metabolomics approaches that use NMR spectroscopy have been used to characterise metabolic variations in response to physiological alternation, disease states, genetic modification, and nutrition intake (Nicholson et al., 1999, 2002). NMR spectroscopy is efficient and cost effective because the analysis is either noninvasive or minimally invasive and requires little sample preparation (Lindon, 2004). The metabolic spectrum from high-resolution NMR spectroscopy usually involves tens of thousands of metabolite features whose intensity values are generated by the resonance of molecules in the sample. Often one wishes to compare a set of spectra from different subjects, conditions, or time points. Such combinations of multiple samples, each with tens of thousands of features, lead to a huge number of data points and a situation that poses a great challenge to analytical capabilities.

A variety of multivariate statistical methods have been introduced to reduce the complexity of metabolic spectra and thus, help identify meaningful patterns in high-resolution NMR spectra (Holmes and Antti, 2002; Lindon et al., 2001). In general, multivariate statistical methods in metabolomics can be divided into two categories, unsupervised and supervised. Principal Components Analysis (PCA) and clustering analysis are examples of unsupervised methods that have been widely used to facilitate the extraction of implicit patterns and elicit the natural groupings of the spectral dataset without prior information about the sample class (Jensen et al., 2004; Beckonert et al., 2003; Solanky et al., 2003). Supervised methods have been applied to classify metabolic profiles according to their various conditions (e.g., time, disease-induced stress, toxic stress, nutritional intake, etc.) (Wang et al., 2004; Beckonert et al., 2003; Holmes et al., 2001; Bathen et al., 2000). The widely used supervised methods in metabolomics include Partial Least Square (PLS) methods, *k*-nearest neighbours, neural networks, and Fisher discrimination analysis. A comprehensive summary of multivariate statistical methods in metabolomics can be found in Holmes and Antti (2002) and Lindon et al. (2001).

In NMR spectra the number of metabolite features present usually greatly exceeds the number of samples, which leads to ill-posed problems. Resolving this difficulty requires an efficient method that can reduce the high dimensionality present to fewer characteristic dimensions that retain most of information of the original data. Although supervised and unsupervised methods have been successfully used for descriptive and predictive analyses in metabolomics, relatively few attempts have been made to identify the metabolite features that play an important role in discriminating between spectra among experimental conditions. The widely used methods include PCA (Goodarcre et al., 2003; Wang et al., 2003) and PLS (Tapp et al., 2003) that provide transformed variables and generally, the first few transformed variables are sufficient to account for the majority variations (e.g., PCA) or

to maximise separability (e.g., PLS) of the entire data. However, extracting meaningful information of original metabolite features from these transformed variables is complicated because these are linear combinations of a large number of original features. Furthermore, selected variables from PCA may not always produce maximum discrimination between classes because PCA does not take into account the class information. Recently, a two-stage genetic programming was introduced for feature selection in metabolomics (Davis et al., 2006). This method obviously compensates for the interpretation problems in PCA and PLS and identifies individual metabolite features necessary for classification. However, genetic programming often involves many parameters that may prevent us from obtaining robust results.

This study presents three feature selection approaches that overcome limitation posed by the transformed variables in PCA and PLS. Three approaches can be categorised into unsupervised, supervised, and a mix of unsupervised and supervised. First, we present the PCA loading method as an unsupervised approach. PCA loadings can be obtained by decomposing transformed variables into components (i.e., loadings) that reflect the contribution of each individual feature. This is a purely unsupervised approach because the process of feature selection is performed without using designated labels of samples. Second, we present an approach that combines unsupervised and supervised feature selection processes. Fisher Discrimination Analysis (FDA) is applied in a reduced dimensional space obtained from PCA. The derived weight vectors obtained through FDA characterise an optimal discriminant direction and the components in weight vectors determine important features for classification. Finally, we use the Variable Importance in Projection (VIP) values from Partial Least Square Discriminant Analysis (PLS-DA) that describes a quantitative estimation of the discriminatory power of each individual feature. This method uses class labels of sample and identifies important individual features that maximise an ability of classification, thus it is a supervised approach. Although the concept of PCA loadings, FDA weight vectors, and PLS-DA VIP values was already introduced in multivariate statistical analysis, their application to metabolomics is still premature. In particular, our study conducts a comparison study of three approaches through real NMR spectra to demonstrate the potential problem of using unsupervised feature selection approaches, widely used in metabolomics.

In addition, we examine the feasibility of using the Orthogonal Signal Correction (OSC) technique along with feature selection to determine whether classification and visualisation could be improved. OSC is a preprocessing technique that removes unwanted spectral variations of data that do not contribute to prediction or classification (Wold et al., 1998). The presented feature selection methods and OSC technique are illustrated using real NMR spectra in which the analytical objective is to identify the metabolite features that characterise metabolic patterns in response to Sulfur Amino Acids (SAA) intake in human plasma. SAA are highly variable in human food, and deficiency and excess are both risks. They are required for physiologic processes in addition to their role in the maintenance of protein synthesis and nitrogen balance.

The remainder of this paper is organised as follows. Section 2 briefly describes the background of PCA, PLS, and OSC necessary for understanding the feature selection approaches presented in Section 3. Section 3 gives detailed descriptions of feature selection approaches. Section 4 describes the experimental data and preprocessing procedures. Section 5 presents the results of the analysis. Finally, Section 6 contains the concluding remarks.

2 Background

2.1 Principal Component Analysis (PCA)

PCA is one of the most frequently used multivariate techniques for dimension reduction. It defines lower-dimensional subspaces that capture as much variation of data matrix **X** ($n \times N$) as possible, where n and N denote the number of samples and features, respectively (Johnson and Wichern, 2002). Mathematically, PCA relies on an eigenvector decomposition of the covariance matrix of **X**, $\text{cov}(\mathbf{X}) = \mathbf{X}^T\mathbf{X}/n - 1$. If only the first A ($A < N$) dimension of scores is needed, PCA decomposes **X** into the sum of the outer products of score vectors (or principal components) $\mathbf{t}_i$ and loading vectors $\mathbf{p}_i$ plus a residual matrix **E**.

$$\mathbf{X}=\mathbf{t}_1\mathbf{p}_1^T+\mathbf{t}_2\mathbf{p}_2^T+\ldots+\mathbf{t}_A\mathbf{p}_A^T+\mathbf{E}=\mathbf{T}\mathbf{P}^T+\mathbf{E}. \quad (1)$$

Here loading vectors $\mathbf{p}_i$ obtained from the eigenvectors of the covariance of **X** account for the contribution of individual features in each principal component dimension (Qin, 2003). The obtained Principal Components (PCs) are uncorrelated with each other, and in general, the first few PCs suffice to characterise the patterns of the spectra.

2.2 Partial Least Squares (PLS)

PLS is a multivariate projection method for modelling a relationship between independent variables **X** and dependent variable(s) **Y**. PLS has been used in various disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science where both the independent and dependent variables are available (Blanco et al., 2000; Shao et al., 2004; Kourti, 2005). PLS seeks to find a set of latent features that maximises the covariance between **X** ($n \times N$) and **Y** ($n \times M$). It decomposes **X** and **Y** into the following forms (Hoskuldsson, 1988):

$$\mathbf{X}=\mathbf{T}\mathbf{P}^T+\mathbf{E}, \quad (2)$$

$$\boldsymbol{Y}=\mathbf{U}\mathbf{Q}^T+\mathbf{F}, \quad (3)$$

where **T** and **U** are ($n \times A$) matrices of the extracted A score vectors, **P** ($N \times A$) and **Q** ($M \times A$) loading matrices, and **E** ($n \times N$) and **F** ($n \times M$) residual matrices. The PLS method searches for weight vectors **w** and **c** that maximises the sample covariance between **t** and **u**. By regressing **X** (**Y**) on **t** (**u**), a loading vector **p** (**q**) can be computed as follows:

$$\mathbf{p}=(\mathbf{t}^T\mathbf{t})^{-1}\mathbf{X}^T\mathbf{t}, \quad (4)$$

$$\mathbf{q}=(\mathbf{u}^T\mathbf{u})^{-1}\boldsymbol{Y}^T\mathbf{u}. \quad (5)$$

Then, the PLS regression model can be expressed as **Y = XB + G**, where **B** and **G** represent regression coefficients and a residual matrix, respectively.

2.3 Orthogonal Signal Correction (OSC)

OSC is a preprocessing technique for removing undesirable systematic variation in data. It was first developed in Wold et al. (1998) to remove systematic variation from the predictor **X** that is orthogonal (or unrelated) to the response **Y**. The largest variation of **X** having zero correlation with **Y** is selectively removed from **X**. The first step of OSC is to calculate the

first PC score vector **t** from **X**. The score vector **t** is then orthogonalised with respect to **Y,** producing the following actual correction vector **t***:

$$\mathbf{t}* = \{\mathbf{I} - \boldsymbol{Y}(\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T\}\mathbf{t}. \quad (6)$$

Then PLS weight vector **w** is computed such that **Xw** = **t***, followed by the calculation of a new score vector, **t** = **Xw**. These processes are repeated until **t** has converged. Finally, a loading vector, **p**, is computed, and the correction term $\mathbf{tp}^T$ is subtracted from **X** giving a residual. The next components can be calculated in such a way. Since the introduction of OSC by Wold et al. (1998), several modified OSC algorithms have been reported (Sjoblom et al., 1998; Fearn, 2000; Westerhuis et al., 2001). In the present study we used a direct orthogonal signal correction algorithm and implemented using MATLAB codes available from Westerhuis et al. (2001). It should be noted that there is a risk of overfitting when too many OSC components are used. In this paper we used two OSC components because previous studies (Wold et al., 1998; Westerhuis et al., 2001) and our own analysis from cross validation indicated that one or two OSC components are sufficient.

3 Feature selection approaches

3.1 PCA loading

Each individual feature does not have the same degree of importance in defining a PCA model. In general, PC, $\mathbf{t}_i$ is the linear combination of the original features weighted by PCA loading coefficients:

$$\mathbf{t}_t = \mathbf{x}_1 p_{t1} + \mathbf{x}_2 p_{t2} + \ldots + \mathbf{x}_N p_{tN} = \mathbf{X}\mathbf{p}_t, \quad i = 1, 2, \ldots, A. \quad (7)$$

The PCA loading coefficients represent the importance of each individual feature in a reduced dimension. For example, p_{i2} indicates the degree of importance of the second original feature in the *i*th PC dimension. In general, two-dimensional loading plots (e.g., p_1–p_2 loading plot) provide useful information to identify important features in the first and second PC dimensions. However, such a use of PCA loading coefficients for feature selection can be extended into *A* PC of interest. A PCA loading-based feature selection index for the *j*th original feature $\mathrm{FSI}_j^{\mathrm{PCA}}$ is given by

$$\mathrm{FSI}_j^{\mathrm{PCA}} = \sum_{i=1}^{A} |p_{ij}| EV_i, \quad j = 1, 2, \ldots, N, \quad (8)$$

where *N* is the total number of metabolite features and EV_i represents the proportion of total variance explained by the *i*th PC.

3.2 FDA weights

FDA is a widely used technique for achieving optimal dimension reduction in classification. FDA provides an efficient lower-dimensional representation of **X** for discrimination among groups of data (Chiang et al., 2000). In other words, FDA uses the dependent variable to seek directions that are optimal for discrimination. This process is achieved by maximising the between-group-scatter matrix $\mathbf{S}_b$ (see equation (9)) while minimising the within-group-scatter matrix $\mathbf{S}_w$ (see equation (9)) (Yang et al., 2004). Thus, FDA finds optimal discriminant weight vectors φ by maximising the following Fisher criterion:

$$J(\boldsymbol{\phi}) = \frac{\boldsymbol{\phi}^T \mathbf{S}_b \boldsymbol{\phi}}{\boldsymbol{\phi}^T \mathbf{S}_w \boldsymbol{\phi}}. \quad (9)$$

It can be shown that this maximisation problem can be reduced to a generalised eigenvalue problem: $\mathbf{S}_b\varphi = \lambda\, \mathbf{S}_w\varphi$ (Yang et al., 2004). The main idea of using FDA for feature selection is to use its weight vector (i.e., $\varphi = [\varphi_1, \varphi_2, \ldots, \varphi_N]^T$). It has been known that feature selection by FDA may encounter computational difficulty due to the singularity of the scatter matrix when the number of samples is smaller than the number of features (Chen et al., 2000). To address this problem, PCA was applied to reduce the number of features, and resulting PC are used as the input features of an FDA classifier. Thus, an FDA weight-based feature selection index for the *j*th original feature $\mathrm{FSI}_j^{\mathrm{FDA}}$ is given by

$$\mathrm{FSI}_j^{\mathrm{FDA}} = \sum_{i=1}^{A} (|p_{ij}| EV_i)\varphi_i, \quad j = 1, 2, \ldots, N. \quad (10)$$

3.3 VIP in PLS-DA

PLS-DA is a special form of PLS for a classification purpose, which explains maximum separation between defined classes of samples. PLS-DA is performed by a PLS regression against a dummy matrix **Y** that indicates class membership (Barker and Rayens, 2003). Each sample is assigned a value of 1 or 0 depending on whether or not it belongs to a specific class. The statistical information obtained from this PLS-DA model can be used to determine which features of **X** are important in determining class membership of **Y** (Musumarra et al., 2004). For this purpose a VIP-based feature selection index for the *j*th original feature, $\mathrm{FSI}_j^{\mathrm{VIP}}$ is computed as

$$\mathrm{FSI}_j^{\mathrm{VIP}} = \frac{N \sum_{i=1}^{A} w_{ij}^2 \mathrm{RSS}_i}{\mathrm{RSS}_T}, \quad j = 1, 2, \ldots, N, \quad (11)$$

where w_{ij} is a PLS-DA weight, RSS_i a percentage of the explained residual sum of squares, and RSS_T a total percentage of the explained residual sum of squares (Kourti and MacGregor, 1996).

4 Experimental data

4.1 Sample collection

We used plasma samples obtained from four healthy subjects under controlled metabolic conditions in the Emory General Clinical Research Center (GCRC). The subjects signed an informed consent approved by the Emory Institutional Review Board and were screened prior to admission with a physician-performed medical history and physical examination, plasma chemistry profile, complete blood count and urinalysis. During the 12-day GCRC admission, the subjects consumed defined diets at standardised intervals. For the first two days (equilibration), the subjects consumed a balanced meal plan with foods selected to ensure adequate energy, protein and SAA intake (SAA at 19 mg/kg/day). After this phase, subjects were placed on constant semi-purified diets designed to alter SAA intake. The diets provided adequate energy and amino acid nitrogen to meet estimated maintenance needs of individual subjects. The semi-purified diet was provided in the form of cookies and

beverages containing L-amino acids, sherbert, corn oil, butter, sugar, and corn starch prepared in the GCRC metabolic kitchen. Daily micronutrient needs were provided in the form of standardised oral doses of multivitamin-mineral supplements, choline, sodium chloride, potassium, and magnesium. The L-amino acid component of the diet was altered to provide zero sulphur amino acids during the initial five days and 117 mg/kg/day during the latter five days of the GCRC stay. Blood was drawn serially 34 times from four subjects over 10 days and proton (^{1}H)-NMR spectra were obtained by a Varian INOVA 600 MHz instrument, which is a high-resolution nuclear NMR spectrometer and used to obtain NMR spectra from blood samples taken serially. During the first 17 time points, blood was collected from subjects consuming zero SAA and 117 mg/kg/day SAA during the latter 17 time points. Figure 1 shows the data structure used for the analysis.

4.2 Preprocessing of NMR spectra

Spectral data generated by ^{1}H-NMR spectroscopy require preprocessing steps *prior to* the subsequent analyses in order to ensure the comparability of multiple spectra. Generally, preprocessing includes phase and baseline corrections, spectra alignment, elimination of redundant regions, and normalisation. Phase and baseline corrections were done using NUTS software (Acorn NMR Inc., Livermore, CA). Variations in spectra caused by concentration, pH, and temperature affect the spectra alignment and thus can interfere with direct comparison between samples. We used a beam search algorithm (Lee and Woodruff, 2004), which determines the best alignment between the spectra by maximising their correlation. Further, redundant regions (e.g., water and the regions containing no metabolite signals) were removed.

A spectrum after removal of the redundant regions is shown in Figure 2(a). Finally, a spectrum was segmented into 0.01 ppm chemical shift bins. The NMR spectrum was reduced to 574 regions (i.e., bins) of equal width (0.01 ppm). The spectral area within bin was integrated using MATLAB (MathWork Inc., Natick, MA). The reduced (binned) spectrum is displayed in Figure 2(b), showing that it retains most of the spectral information contained in the original spectrum.

5 Results

5.1 Effect of OSC

Before undertaking the feature selection process, OSC was performed to improve the separability of the two different dietary phases (zero SAA and supplemented SAA) by removing unwanted variation that does not contribute to discrimination. Figure 3(a) shows PCA score plots from the NMR spectra not processed by OSC. The PC1 and PC2 components, respectively, in Figure 3(a) explained 68.1% and 14.1% of the total variation of **X**.

Figure 3(b) and (c) show PCA and kernel PCA (KPCA) score plots obtained from OSC-processed spectra. We attempt to use KPCA, expecting that visualisation is improved by dealing with the nonlinear characteristics of spectra. The basic idea of KPCA is to first map input data into a nonlinear feature space F and then to extract PC in that feature space. Replacing canonical dot products in F by a kernel function (e.g., Gaussian, polynomial, and sigmoid functions) eliminates the need to execute nonlinear mappings and dot products in F (Cortes and Vapnik, 1995). To our best knowledge, this is the first work that attempts to adopt KPCA in metabolomics.

The two-dimensional score plots of PCA and KPCA using the OSC processed data clearly show better separation between the zero-SAA phase and the SAA-supplemented phase (Figure 3(b) and (c)) compared with the results without OSC (Figure 3(a)). In OSC-KPCA, a

Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/c)$ was chosen with $c = 6$ because it yielded better separation between the two dietary phases than other kernel functions and parameter values. The use of the kernel function is to represent the nonlinearity of the NMR data and to find out the separation between the two dietary phases in reduced score spaces. The OSC-KPCA score plot seems to provide slightly better separation than OSC-PCA. Overall, clearer separation between the two dietary phases was achieved in linear (OSC-PCA) and nonlinear (OSCKPCA) PC reduced spaces processed by OSC. This most likely leads to less misclassification rate in classification models discussed in later section. The PCA, OSC-PCA, and OSC-KPCA models give an overall impression of how well the different classes can be separated. The subsequent classification analysis of the NMR spectral data was performed with the two dietary phases. Original features and PCs were used for calculating classification results based on the *k*-Nearest Neighbour (*k*-NN) method. The experimental data were split into four groups corresponding to four individuals; each group includes 17 spectra obtained during the zero-SAA phase and 17 samples obtained during the supplemented-SAA phase. Three groups of the samples were used for training, and the one remaining group was kept aside as testing data. This process was repeated three times more to obtain an overall misclassification rate.

The classification results from the four different sets of testing were averaged (Table 1). The best classification was achieved using OSC-PCA and OSC-KPCA, both of which yielded a zero misclassification rate in both the training and testing samples. On the other hand, the two classification methods not processed by OSC produced relatively less accurate classification results. In *k*-NN, different values of neighbourhood parameters *k* were examined to determine which has the lowest misclassification rate (minimum error at $k = 4$). In addition, two PLS-DA models are built to investigate the OSC filtering effect, the results of which are shown in Table 2. It turned out that the “OSC-processed PLS-DA” model has better predictive power (i.e., discriminative, in this case) than the ‘PLS-DA’ model: 0.886 vs. 0.225. Such an improvement in discrimination can be also shown in score plots of the two PLS-DA models (Figure 4). The “OSC-processed PLS-DA” (Figure 4(b)) shows two distinct clusters that represent a well-defined discrimination of the two classes of dietary SAA intake.

5.2 Feature selection and evaluation

The three feature selection approaches described in Section 3 were performed to find important metabolite features that contribute to the discrimination between the zero-SAA phase and the supplemented-SAA phase. For feature selection we used OSC-processed spectra because they gave a well-defined discrimination of the NMR spectra with two distinct classes. PCA loading values ($\mathrm{FSI}_j^{\mathrm{PCA}}$ for $j = 1, \ldots, N$), FDA weights ($\mathrm{FSI}_j^{\mathrm{FDA}}$ for $j = 1, \ldots, N$), PLS-DA VIP scores ($\mathrm{FSI}_j^{\mathrm{VIP}}$ for $j = 1, \ldots, N$) for individual metabolite features were obtained using equations (8), (10) and (11), respectively and they are plotted against chemical shifts (Figure 5). To calculate PCA loading values, we used seven PCs, which explain 96.5% of total variation of the entire data.

The calculation of $\mathrm{FSI}_j^{\mathrm{VIP}}$ allows us to rank each of the 574 individual features according to its contribution to the capability to discriminate different classes of the experimental data. A total of 269 metabolite features with large VIP values (i.e., VIP > 1) were selected (see Figure 5(c)). The features with VIP > 1 were considered to be important because the squared sum of all VIP values is equal to the number feature and thus, the average VIP would be equal to 1 (Umetrics, 2005).

To ensure the comparability of the three feature selection methods, the same number of features was selected for both the PCA loading approach and the FDA weight approach. The threshold values to determine important features were 0.03332 for the PCA loading method (see Figure 5(a)) and 3.08×10^{-4} for the FDA weight method (see Figure 5(b)). It can be seen that the similar metabolite features were selected from the PCA loading and FDA weight approaches because both approaches adopted an unsupervised manner. PLS-DA VIP that uses a supervised feature selection process identified a different set of metabolite features compared to other two approaches. More specifically, most of metabolite features between 0.53 ppm and 2.53 ppm were selected as significant from PLS-DA VIP, while this was not the case in PCA loading and FDA weight. Furthermore, it is interesting to note that the PCA loading and FDA weight approaches identified all the metabolite features between 6.8 ppm and 7.8 ppm as significant, while none of them were selected as important from the PLS-DA VIP approach.

To evaluate a set of metabolite features selected by three feature selection approaches, classification models were developed with those features. The *k*-NN models with $k = 4$ were developed with the experimental data. As before, the experimental data were split into four groups corresponding to the four individuals. Three individuals were used for training the models, and the one remaining individual was used for testing. This process was repeated three more times to obtain the cross-validated error rate (misclassification rate). Misclassification rates from three approaches were summarised in Table 3. The best classification was achieved using the VIP values of the PLS-DA model: Average misclassification rates are 0% in the training data and 0.7% in the testing data. This result implies that an unguarded use of unsupervised feature selection approaches may mislead features selection results.

6 Conclusions

Systematic application of the features selection approaches to metabolomics can provide a basis for simultaneous determinations of multiple nutritional endpoints and the modelling and analysis using individual features in complex high-resolution NMR spectra.

We have presented three feature selection approaches (PCA loadings, FDA weight vectors, and PLS VIP values) for high-resolution NMR spectra to identify informative metabolite features in human plasma that characterise metabolic perturbations induced by dietary SAA intake. These approaches can offer advantages over conventional PCA and PLS techniques because they take into account the decomposition of transformed variables that provide a clear interpretation with respect to the original metabolite features. Furthermore, the effect of using OSC as a preprocessing step has been illustrated in the PCA and the KPCA score plots by showing significant improvement in the separation of samples. This led to more accurate classification results. The overall result has shown that the better classification was achieved using VIP values from an OSC-processed PLS-DA model than other two approaches using unsupervised manners. This result implies that a mere use of unsupervised features approaches in labeled data, widely used in metabolomics for feature selection, lead to the potential problems. We hope that the approaches and comparison results presented in this paper stimulate further investigation in the development of better analytic approaches for feature selection in metabolomics.

Acknowledgments

The authors would like to thank the editor and three anonymous reviewers whose comments helped significantly improve the quality of this paper. We are grateful to nursing and laboratory staff of the Emory General Clinical Research Center for valuable helps in collecting samples. This study was supported by grants from the National

Institutes of Health: R03 DK066008, R03 ES012929, R01 ES011195, R01 DK55850, and the Emory General Clinical Research Center grant M01 RR00039.

Biographies

Biographical notes: Hyun-Woo Cho is a post-doctoral research associate in the Department of Industrial and Information Engineering at the University of Tennessee at Knoxville. He received his PhD in Industrial Engineering in 2003 from POSTECH in South Korea and MS/BS in Chemical Engineering. He is a member of the Institute for Operations Research and Management Sciences (INFORMS). His research interests include statistical pattern recognition, NMR/NIR-based feature selection, and nonlinear kernel/wavelet-based analysis for manufacturing, chemical, bio, and bioenergy processes.

Seoung Bum Kim is an Assistant Professor of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. He was a post-doctoral fellow at the Emory University Medical School. He received his PhD in Industrial and Systems Engineering in 2005 from the Georgia Institute of Technology. He was awarded the Jack Youden Prize as the best expository paper in Technometrics for the Year 2003. He is a member of the Institute for Mathematical Statistics. His research interests include statistical and data mining modelling of high-resolution NMR spectra.

Myong K. Jeong is currently an Assistant Professor in the Department of Industrial and Information Engineering, the University of Tennessee, Knoxville. He received the PhD Degree in Industrial and Systems Engineering from the Georgia Institute of Technology, Atlanta, Georgia, in 2004. His research interests include data mining, pattern recognition, NMR/NIR spectral analysis, and wavelets. He is a member of IEEE and INFORMS. He won the Freund International scholarship in 2002.

Youngja Park, PhD is an instructor in the Department of Medicine (Pulmonary, Allergy and Critical Care Division) at Emory University, Atlanta, GA. She received a PhD in Pharmacology and Toxicology from University of Texas at Austin, in 1990. Her central research focus is on nutritional metabolomics using ^{1}H-NMR. She is currently working in the Emory Clinical Biomarkers Laboratory and the NMR Research Center to collect NMR spectral data of human plasma. She has developed procedures for management and statistical analysis of NMR spectra of human plasma.

Nana Gletsu is an instructor in the Department of Surgery (General and Gastrointestinal Surgery at Emory University, Atlanta, GA. She received a PhD in Nutrition and Metabolism at University of Alberta, Edmonton, Canada in 1998. Her research interests include understanding the links between obesity and oxidative stress/insulin resistance/metabolic syndrome. Her present studies include identifying biomarkers indicative of insulin resistance using proteomics and metabolomics.

Dean P. Jones, PhD is a Professor in the Department of Medicine (Pulmonary, Allergy and Critical Care Division) at Emory University, Atlanta, GA. He received a PhD in Biochemistry from Oregon Health Sciences Univ., Portland, in 1976. His central research focus is on redox mechanisms of oxidative stress. He currently directs the Emory Clinical Biomarkers Laboratory, which is focused on oxidative stress biomarkers and applications of ^{1}H-NMR spectroscopy and Fourier-transform mass spectrometry for high-throughput clinical metabolomic analyses of nutritional and environmental factors in human health and disease.

Thomas R. Zeigler is an Associate Professor in the Department of Medicine (Endocrinology, Metabolism and Lipids Division) at Emory University, Atlanta, GA. He received his MD Degree from the Michigan State University College of Human Medicine, E. Lansing, MI, in 1983. He is currently Associate Director of the Emory General Clinical Research Center. His central research focus is in the area of enteral and parenteral nutrition support and nutrient-growth factor interactions in clinical and translational models of malnutrition and catabolic stress.

References

Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics. 2003; 17:166–173.

Bathen TF, Krane J, Engan T, Bjerve KS, Axelson D. Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis. NMR Biomed. 2000; 13:271–288. [PubMed: 10960918]

Beckonert O, Bollard ME, Ebbels TMD, Keun HC, Antti H, Holmes E, Lindon JC, Nicholson JK. NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. Analytica Chimica Acta. 2003; 490:3–15.

Blanco M, Coello J, Eustaquio A, Iturriaga H, Maspoch S. Development and validation of a method for the analysis of a pharmaceutical preparation by near-infrared diffuse reflectance spectroscopy. Journal of Pharmaceutical Sciences. 2000; 88:551–556. [PubMed: 10229648]

Chen L-F, Lio H-YM, Ko M-T, Lin J-C, Yu G-J. A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition. 2000; 33:1713–1726.

Chiang LH, Russell EL, Braatz RD. Fault diagnosis in chemical processes using Fisher discriminant analysis, discriminant partial least squares, and principal component analysis. Chemometrics and Intelligent Laboratory Systems. 2000; 50:243–252.

Cortes C, Vapnik VN. Support vector networks. Machine Learning. 1995; 20:273–297.

Davis RA, Charlton AJ, Oehlschlager S, Wilson JC. Novel feature selection method for genetic programming using metabolomic ^{1}H NMR data. Chemometrics and Intelligent Laboratory Systems. 2006; 81:50–59.

Fearn T. On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2000; 50:47–52.

Goodarcre R, York EV, Heald JK, Scott IM. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. Phytochemistry. 2003; 62:859–863. [PubMed: 12590113]

Holmes E, Antti H. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst. 2002; 127:1549–1557. [PubMed: 12537357]

Holmes E, Nicholson JK, Tranter G. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. Chemical Research in Toxicology. 2001; 14:182–191. [PubMed: 11258967]

Hoskuldsson A. PLS regression methods. Journal of Chemometrics. 1998; 2:211–228.

Jensen JJ, Hoefsloot HCJ, Boelens HFM, Greef JVD, Smilde AK. Analysis of longitudinal metabolomics data. Bioinformatics. 2004; 20:2438–2446. [PubMed: 15087313]

Johnson, RA.; Wichern, DW. Applied Multivariate Statistical Analysis. Prentice-Hall; New York: 2002.

Kourti T. Application of latent variable methods to process control and multivariate statistical process control in industry. International Journal of Adaptive Control and Signal Processing. 2005; 19:213–246.

Kourti T, MacGregor JF. Multivariate SPC methods for process and product monitoring. Journal of Quality Technology. 1996; 28:409–428.

Leardi R. Genetic algorithms in chemometrics and chemistry: a review. Journal of Chemometrics. 2001; 15:559–569.

Lee GC, Woodruff DL. Beam search for peak alignment of NMR signals. Analytica Chimica Acta. 2004; 513:413–416.

Lindon JC. Metabonomics – techniques and applications. Business Briefing: Future Drug Discovery. 2004; 2004:1–6.

Lindon JC, Holmes E, Nicholson JK. Pattern recognition methods and applications in biomedical magnetic resonance. Progress in Nuclear Magnetic Resonance Spectroscopy. 2001; 39:1–40.

Musumarra G, Barresi V, Condorelli DF, Fortuna CG, Scire S. Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysis. Journal of Chemometrics. 2004; 18:125–132.

Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. Nature Reviews Drug Discovery. 2002; 1:153–161.

Nicholson JK, Lindon JC, Holms E. Metabonomics: Understanding the metabolic responses of living systems to pathophysiological stimuli via multi-variate statistical analysis of biological NMR spectroscopic data. Xenobiotica. 1999; 29:1181–1189. [PubMed: 10598751]

Qin SJ. Statistical process monitoring: basics and beyond. Journal of Chemometrics. 2003; 17:480–502.

Shao X, Wang F, Chen D, Su Q. A method for near-infrared spectral calibration of complex plant samples with wavelet transform and elimination of uninformative variables. Analytical and Bioanalytical Chemistry. 2004; 378:1382–1387. [PubMed: 14735278]

Sjoblom J, Svensson O, Josefson M, Kullberg H, Wold S. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1998; 44:229–244.

Solanky KS, Bailey NJC, Beckwith-Hall BM, Davis A, Bingham S, Holmes E, Nicholson JK, Cassidy A. Application of biofluid ^{1}H nuclear magnetic resonance-based metabonomic technique for the analysis of the biochemical effects of dietary isoflavones on human plasma profile. Analytical Biochemistry. 2003; 323:197–204. [PubMed: 14656525]

Sun J. Statistical analysis of NIR data: data pretreatment. Journal of Chemometrics. 1997; 11:525–532.

Tapp HS, Defernez M, Kemsley EK. FTIR spectroscopy and multivariate analysis can distinguish geographical origin of extra virgin olive oil. Journal of Agricultural and Food Chemistry. 2003; 51:6110–6115. [PubMed: 14518931]

Umetrics. SIMCA-P and SIMCA-P+ User Guide and Tutorial. 2005.

Wang YL, Bollard ME, Keun H, Antti H, Beckonert O, Ebbels TM, Lindon JC, Holmes E, Tang HR, Nicholson JK. Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning H-1 nuclear magnetic resonance spectroscopy of liver tissues. Analytical Biochemistry. 2003; 322:26–32. [PubMed: 14705776]

Wang YL, Holmes E, Nicholson JK, Cloarec O, Chollet J, Tanner M, Singer BH, Utzinger J. Metabonomic investigations in mice infected with Schistosoma mansoni: an approach for biomarker identification. Proc Natl Acad Sci USA. 2004; 101:12676–12681. [PubMed: 15314235]

Westerhuis JA, de Jong S, Smilde AK. Direct orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2001; 56:13–25.

Wold S, Antti H, Lindgren F, Öhman J. Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems. 1998; 44:175–185.

Yang J, Frangia AF, Yang J. A new kernel Fisher discriminant algorithm with application to face recognition. Neurocomputing. 2004; 56:415–421.

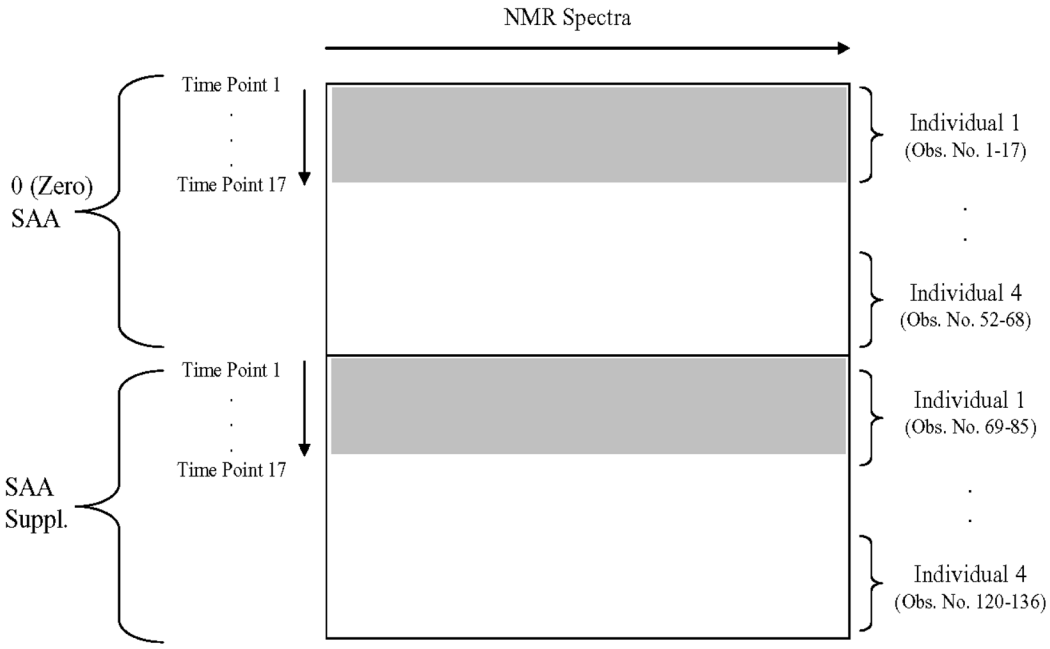

Figure 1.
A schematic diagram of data structure used in the experimental study

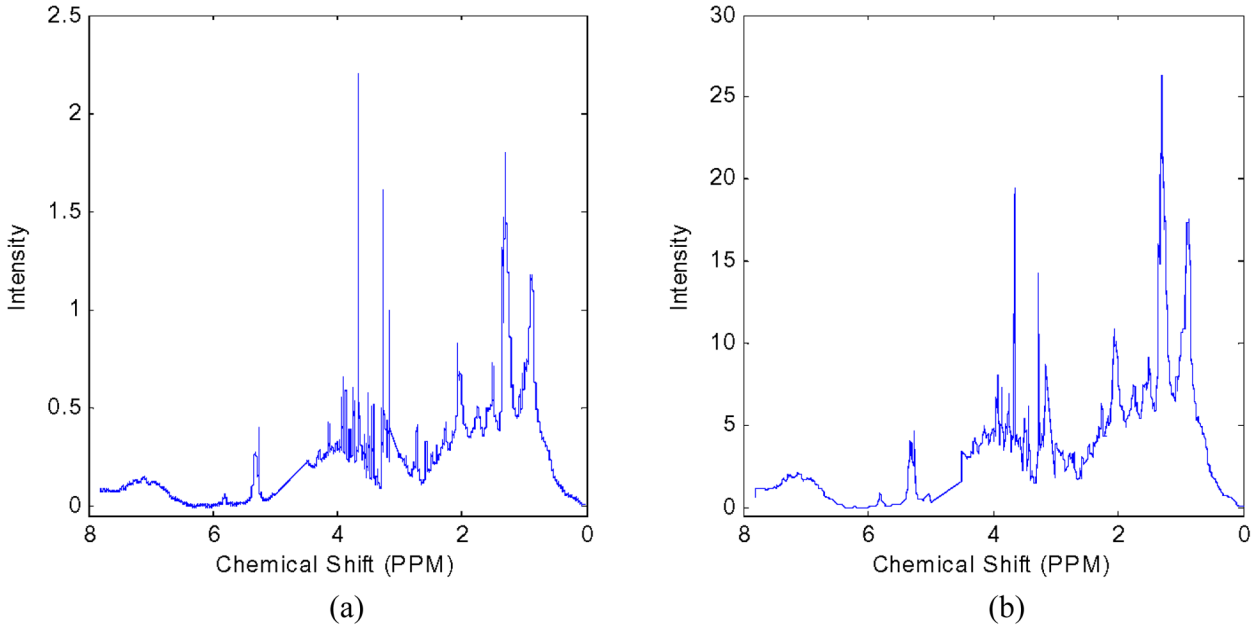

Figure 2.
^{1}H NMR spectral data for (a) original (number of metabolite features = 8,445) and (b) reduced (binned) spectrum (number of metabolite features = 574)

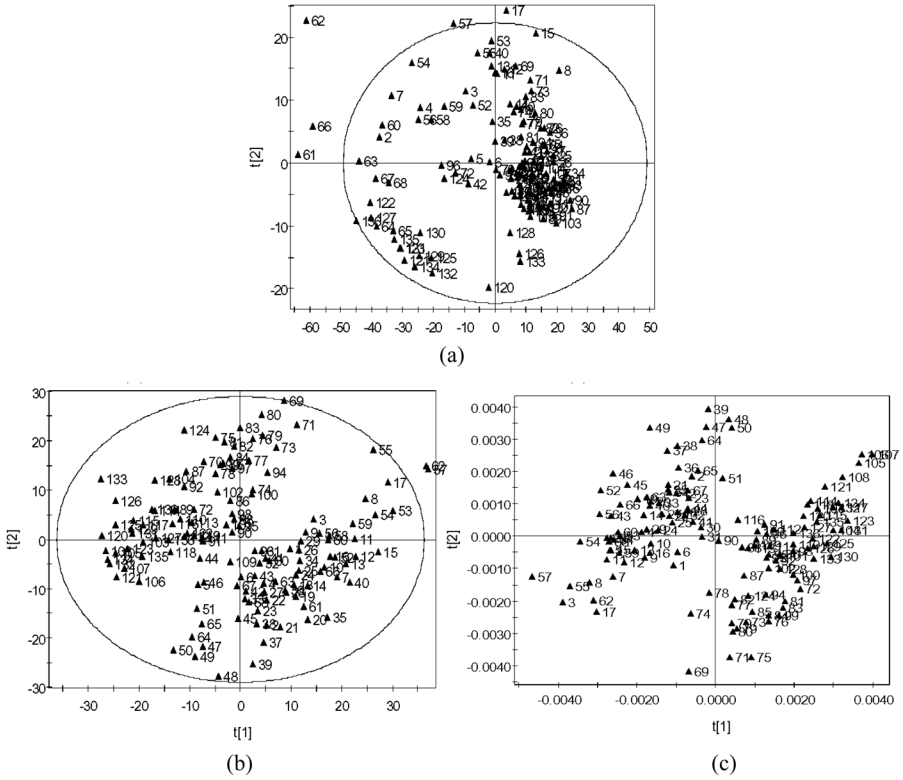

Figure 3.
Score plots obtained from (a) PCA; (b) OSC-processed PCA and (c) OSC-processed KPCA

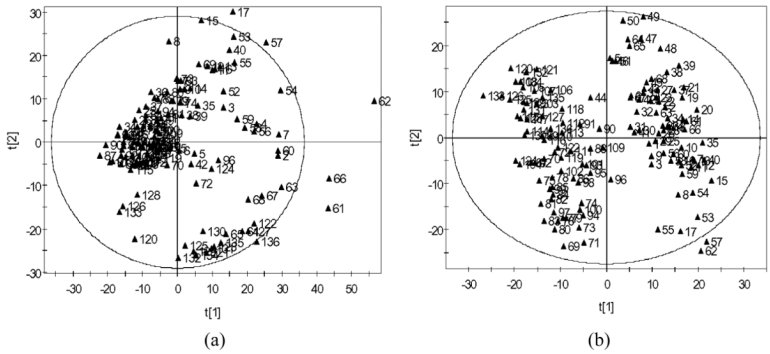

Figure 4.
Score plots for (a) PLS-DA and (b) OSC-processed PLS-DA models

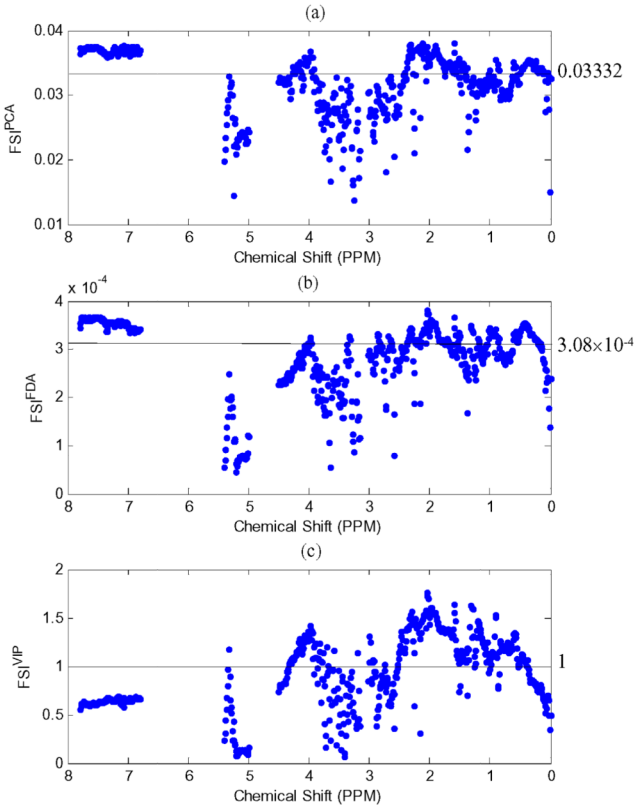

Figure 5.
Feature selection results based on (a) OSC-PCA loading; (b) SC-PCA loading with FDA weights and (c) VIP of OSC-PLS-DA

Table 1

Comparison of misclassification rates (%) of *k*-NN using original features, PC scores, OSC-PCA PC scores, and OSC-KPCA PC scores

		Original features (574 metabolite features)					
		k = 3	**k = 4**	**k = 5**	**PCA PC scores**	**OSC-PCA PC scores**	**OSC-KPCA PC scores**
Training	Zero SAA	19.1	9.3	28.4	16.2	0.0	0.0
	SAA supplement	10.3	6.9	19.1	11.8	0.0	0.0
	Total	*14.7*	*8.0*	*16.4*	*14.0*	*0.0*	*0.0*
Testing	Zero SAA	41.1	36.8	44.1	30.9	0.0	0.0
	SAA supplement	45.6	33.8	27.9	57.4	0.0	0.0
	Total	*43.4*	*35.3*	*36.0*	*44.0*	*0.0*	*0.0*

Table 2

The performance of PLS-DA models with/without OSC preprocessing

	PLS-DA			*OSC-processed PLS-DA*		
A *	***R*²X (*cum*)**+	***R*²Y (*cum*)**+	***Q*²(*cum*)**+	***R*²X (*cum*)**	***R*²Y (*cum*)**	***Q*² (*cum*)**
1	0.580	0.130	0.120	0.339	0.824	0.822
2	0.830	0.240	0.225	0.596	0.893	0.886

* indicates the number of latent components retained in PLS-DA models.

+ R^2**X**(cum) and R^2**Y** (cum) is a cumulative sum of the squares of **X** and **Y** explained, respectively, and Q^2 (cum) represents a cumulative fraction of the total variance of **Y** predicted by extracted components.

Table 3

Comparison of misclassification rates (%) using a set of metabolite features selected by OSC-PCA loading, OSC-PCA loading with FDA weights, and VIP of OSC-PLS-DA

		OSC-PCA loading	*OSC-PCA loading with FDA weights*	*VIP of OSC-PLS-DA*
Training	Zero SAA	1.5	1.5	0.0
	SAA supplement	0.0	0.0	0.0
	Total	*0.7*	*0.7*	*0.0*
Testing	Zero SAA	2.9	2.9	1.5
	SAA supplement	4.4	4.4	0.0
	Total	*3.7*	*3.7*	*0.7*